2020-12

MULTI-LABEL CLASSIFICATION OF UNDERWATER SOUNDSCAPES USING DEEP CONVOLUTIONAL NEURAL NETWORKS

Pfau, Andrew M.

Monterey, CA; Naval Postgraduate School

http://hdl.handle.net/10945/66705

NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

MULTI-LABEL CLASSIFICATION OF UNDERWATER SOUNDSCAPES USING DEEP CONVOLUTIONAL NEURAL NETWORKS

by

Andrew M. Pfau

December 2020

Thesis Advisor:	Marko Orescanin
Second Reader:	Geoffrey G. Xie

Approved for public release. Distribution is unlimited.

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE	*Form Approved OMB No. 0704-0188*

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington, DC, 20503.

1. AGENCY USE ONLY ***(Leave blank)***	**2. REPORT DATE** December 2020	**3. REPORT TYPE AND DATES COVERED** Master's thesis

4. TITLE AND SUBTITLE MULTI-LABEL CLASSIFICATION OF UNDERWATER SOUNDSCAPES USING DEEP CONVOLUTIONAL NEURAL NETWORKS	**5. FUNDING NUMBERS**
6. AUTHOR(S) Andrew M. Pfau	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943-5000	**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) N/A	**10. SPONSORING / MONITORING AGENCY REPORT NUMBER**

11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release. Distribution is unlimited.	**12b. DISTRIBUTION CODE** A

13. ABSTRACT (maximum 200 words)

The detection and classification of passive sonar acoustics is a challenging problem faced by surface, subsurface, and naval air assets. The potential benefit of machine learning systems to assist in this task is appealing. However, little work has been conducted to develop and test machine learning models for this type of data or task. This thesis presents a custom convolutional neural network (CNN) model designed specifically for underwater acoustic classification. This model is compared to several common CNN architectures on two datasets of hydrophone recordings of passing ships. These datasets are some of the largest datasets of ship recordings used for training CNNs to date, composed of over 4,000 hours of recordings and hundreds of unique ships. This thesis's main contribution is in demonstrating multi-label classification on underwater ship acoustics where the proposed model achieved an average micro-F1 score of 0.97. The custom CNN shows marked improvement in performance over standard models in both multi-class and multi-label classification tasks. This work also presents research into the inclusion of synthetic ship sounds and their potential use in training classification models. This thesis demonstrates the capability of machine learning models to enhance human and unmanned systems operating in the undersea domain.

14. SUBJECT TERMS artificial intelligence, neural networks, soundscape classification, machine learning	**15. NUMBER OF PAGES** 61
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT Unclassified	**18. SECURITY CLASSIFICATION OF THIS PAGE** Unclassified	**19. SECURITY CLASSIFICATION OF ABSTRACT** Unclassified	**20. LIMITATION OF ABSTRACT** UU

THIS PAGE INTENTIONALLY LEFT BLANK

MULTI-LABEL CLASSIFICATION OF UNDERWATER SOUNDSCAPES USING DEEP CONVOLUTIONAL NEURAL NETWORKS

Andrew M. Pfau
Lieutenant, United States Navy
BS, U.S. Naval Academy, 2014

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN COMPUTER SCIENCE

from the

NAVAL POSTGRADUATE SCHOOL
December 2020

Approved by: Marko Orescanin
Advisor

Geoffrey G. Xie
Second Reader

Gurminder Singh
Chair, Department of Computer Science

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

The detection and classification of passive sonar acoustics is a challenging problem faced by surface, subsurface, and naval air assets. The potential benefit of machine learning systems to assist in this task is appealing. However, little work has been conducted to develop and test machine learning models for this type of data or task. This thesis presents a custom convolutional neural network (CNN) model designed specifically for underwater acoustic classification. This model is compared to several common CNN architectures on two datasets of hydrophone recordings of passing ships. These datasets are some of the largest datasets of ship recordings used for training CNNs to date, composed of over 4,000 hours of recordings and hundreds of unique ships. This thesis's main contribution is in demonstrating multi-label classification on underwater ship acoustics where the proposed model achieved an average micro-F1 score of 0.97. The custom CNN shows marked improvement in performance over standard models in both multi-class and multi-label classification tasks. This work also presents research into the inclusion of synthetic ship sounds and their potential use in training classification models. This thesis demonstrates the capability of machine learning models to enhance human and unmanned systems operating in the undersea domain.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

I. INTRODUCTION 1

A. RESEARCH QUESTIONS AND CONTRIBUTIONS 2

B. OUTLINE 3

II. PREVIOUS WORK 5

A. TRADITIONAL MACHINE LEARNING METHODS 5

B. NEURAL NETWORK BASED APPROACHES 6

1. Introduction to Neural Networks 6

2. Convolutional Neural Networks 7

3. CNN Filter Shapes 9

4. Dataset Requirements...10

C. MULTI-CLASS VERSUS MULTI-LABEL...11

III. UNDERWATER SOUNDS 13

A. UNDERWATER SOUNDS 13

B. SHIP ACOUSTICS 14

C. SPECTROGRAM GENERATION...16

D. DATASETS 19

1. Dataset Generation 19

2. Dataset Labeling..20

IV. METHODOLOGY 23

A. CUSTOM CNN ARCHITECTURE...23

B. EXPERIMENT METHODOLOGY 25

C. GAN DATA GENERATION...26

V. RESULTS 27

A. MULTI CLASS 27

1. HARP Dataset 27

2. MARS Dataset...28

3. Vector Sensor 29

B. MULTI-LABEL 30

C. DATA AUGMENTATION WITH GANS..32

1. Evaluation of Synthetic Data Realness 32

2. Augmenting with Synthetic Data..32

3. Boosting Single Class Performance 33

VI. CONCLUSIONS AND FUTURE WORK 35
- A. CONCLUSION 35
- B. FUTURE WORK 36

LIST OF REFERENCES 39

INITIAL DISTRIBUTION LIST 43

LIST OF FIGURES

Figure 1. Diagram of Simplified CNN. Source: [7]. 7

Figure 2. Chart Showing Various Sources of Underwater Sound and Their Frequency. Source: [20]. 14

Figure 3. Chart Showing Relationship between Ship Size, Speed, and Radiated Source Level. Source: [22]. 16

Figure 4. Mel Spaced Filter Bank. Source: [28]...18

Figure 5. Example Spectrograms 18

Figure 6. Classifier Accuracy with Respect to Kernel Ratio 23

Figure 7. Diagram of Proposed Custom CNN Architecture 24

Figure 8. Time-Range Plot of Passing Ship 28

Figure 9. Time-Range Plot Showing Classification when Using Multi-Label Classification...31

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF TABLES

Table 1. Dataset Class List 21

Table 2. Table 2: HARP Multi-class Results 27

Table 3. MARS Multi-class F1 Score Results 29

Table 4. F1 Scores for Multi and Single Channel Data 30

Table 5. Multi-label Dataset Label Distribution 30

Table 6. Multi-label Results 31

Table 7. Accuracy of Classifier When Trained on Real Data Only 32

Table 8. F1 Scores for Training with Synthetic Samples 33

Table 9. F1 Scores for Custom Model for Single Class Boosting 33

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF ACRONYMS AND ABBREVIATIONS

AED	acoustic event detection
AI	artificial intelligence
AIS	Automatic Identification System
CNN	convolutional neural network
DON	Department of the Navy
FFT	fast Fourier transform
GAN	generative adversarial network
HARP	High Frequency Acoustic Recording Package
IMO	International Maritime Organization
ML	machine learning
MLP	multi layer perceptron
MMSI	Maritime Mobile Service Identity
ReLU	rectified linear unit
STFT	short time Fourier transform
SVM	support vector machine
UUV	Unmanned Undersea Vehicle

THIS PAGE INTENTIONALLY LEFT BLANK

ACKNOWLEDGMENTS

I would like to thank my thesis advisor, Professor Orescanin, for all of his assistance and for his shared interest in this topic. I am very grateful to have found an advisor and thesis topic that closely aligned with my background and interests. I would also like to thank Professor Xie for his insightful comments and feedback on my thesis as my second reader. Special thanks to the members of the Oceanography and Physics departments, who generously provided their datasets and assistance in understanding them.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

Proliferation of Artificial Intelligence (AI) technology across all segments of society has accelerated in recent years. Reacting to this proliferation, the Department of the Navy (DON) is focusing on ways to leverage this technology to enhance military capabilities. One of the ways for AI to aid human operators is in making sense of the vast amounts of data that the DON collects every day. AI can automate tasks such as segmentation and classification of data much faster and more accurately than a human analyst can alone. Analysts often must sift through troves of data from remote sensor networks, satellite imagery, or unmanned aircraft. Enhancing human analysts' workflows with AI systems will assist them in more rapidly and better understanding this data.

To accomplish the task of automating extraction of meaningful information from acoustic data most of the current research is focused on the use of supervised learning algorithms, specifically, deep convolutional neural networks (CNNs). The use of CNNs in automatic classification has roots in computer vision but has recently became very popular for soundscape classification [1]. In this approach, commonly recorded sound audio is transformed into time-frequency spectrograms or their derivatives. Before CNNs, other methods of automatic classification of sounds were used with varying degrees of success, including Support Vector Machines (SVM) and Stacked Auto Encoders (SAE) [2], [3]. However, the ability of CNNs to generalize knowledge to a broad set of input examples makes them a better option when dealing with a large set of potential object classes.

There are two areas within the DON that can best leverage this technology, unmanned vehicles and remote sensor networks employed in both surveillance and defensive monitoring.

The Navy has undertaken an effort to standup a fleet of both unmanned surface vessels (USVs) and unmanned underwater vessels (UUVs). In September of 2017, the Navy established Unmanned Undersea Vehicle Squadron 1 to procure, test, and deploy UUVs. The squadron is responsible for UUVs in a range of sizes from 10 inches diameter to the 51 ft long Orca XLUUV, currently under construction [4], [5]. While these UUVs possess increasing capabilities and underwater endurance they lack the ability to appropriately detect and respond to other ships. Onboard a manned submarine, a sonar operator’s task is to “detect, track, and classify” any ship or other submarine that the sonar system detects. This complex task is performed with a combination of visual displays, auditory listening, and operator experience. Mimicking this ability on an unmanned system is challenging given the wide range of sounds and scenarios that a sonar operator can encounter. Designing a system that can accurately detect and classify a range of underwater sounds is the first step in making UUVs truly capable of self-sufficient operations.

Similarly, human analysts are required to monitor remote sensors for threats or intelligence and use both visual and auditory means. Harbor security is one application of remote sensors for both military and civilian ports. This type of security currently employs radar, Automatic Identification System (AIS) monitoring, and some sonar systems. These systems are designed to detect surface intruders and would be unable to detect subsurface intruders such as UUVs or divers [6]. With the advent of cheaper sensors, the ability to automatically monitor a network of sensors can be employed to guard a harbor against intrusion or sabotage of ships in port using passive monitoring sensors such as hydrophone or combination of active and passive monitoring.

A. RESEARCH QUESTIONS AND CONTRIBUTIONS

This work is centered around addressing the challenges of heterogenous underwater soundscape classifications in coastal zones with a focus on classifying shipping noises. In contrast to the common approach of rare acoustic event classification, here the goal is to detect multiple target labels per inference of the neural network classifier. First, single class labels are applied to the dataset and multi-class classification is evaluated. This type of classification is performed on both single and multi-channel data, exploring whether or not additional data gather by the sensor is beneficial in the classification process. Then, multi-label classification is explored, where more than one class of ship is present in the same sample. This is achieved by developing and evaluating a multi-label classification CNN architecture. Finally, the benefit of augmenting the existing dataset with synthetic data is explored through the use of a Generative Adversarial Network (GAN) to generate realistic ship like sounds. The objective of augmenting the dataset with synthetic data is to increase the diversity of class representation especially for the unrepresented classes in the dataset. Specifically, in this work, the objective is to test whether GAN generated sounds can be correctly classified by a model trained on real data.

In order to accomplish the classification tasks, several CNN architectures are adapted, and their performance is compared. A custom CNN architecture is designed to take advantage of the unique time frequency aspects of acoustic data. The proposed architecture is compared against popular CNN architectures designed for image classification to benchmark performance of the proposed architecture. Further, advantages of the proposed architecture for these tasks are demonstrated through rigorous studies capturing multiple hyperparameters of the neural network design as well.

There are several key contributions from this research. This research presents the first instance, to our knowledge, of the multi-label classification on underwater soundscapes reported in literature. It is also the first use, to our knowledge, of multi-channel vector sensor data for soundscape classification using deep learning. Secondary contributions involve curation and evaluation of deep learning models on a large dataset of recordings of hundreds of unique ships, which represents one of the largest labeled databases used in this type of study; the development of a CNN architecture specifically designed for classification of underwater ship acoustics, and the exploration of synthetic data to augment training for underwater acoustic classification tasks utilizing GAN architectures.

B. OUTLINE

Chapter II provides background information on neural networks and machine learning for multi label problems. Chapter III provides on introduction to underwater acoustics, how ships generate noises, and challenges of acoustic classification in the underwater domain. Chapter IV discusses the dataset used in this thesis and the research methodology and CNN architecture design. Results and analysis of results are discussed in Chapter V.

THIS PAGE INTENTIONALLY LEFT BLANK

II. PREVIOUS WORK

Machine learning techniques have been applied to the field of acoustics with attempts to classify specific sounds within a recording, often referred to as acoustic event detection (AED), and acoustic scene classification. In this section we discuss the evolution of these techniques from traditional methods to current deep learning methods.

A. TRADITIONAL MACHINE LEARNING METHODS

Traditional methods of machine learning rely on hand-designed features to extract useful information from the signals of interest. These features are then passed over the audio data looking to match the signal of interest. This is a labor-intensive process that requires expert knowledge in the field of underwater acoustics and a different set of features for each signal of interest. In the audio domain these features take the form of matching the changes in frequency over time of a target of interest. For example, a classifier would require hand-engineered features for each type of whale, dolphin call, or ship type that the system seeks to classify. These features are also hand-engineered and finely tuned to the available example data by the human expert who designs them. Slight changes in the sweep of a dolphin call may not be classified if there is not enough tolerance for error. Conversely too much tolerance will produce too many false positive classifications. Additionally, new signals introduced to the classifier may not be identified because they do not meet the design of the feature extractors.

In [6] researchers classified whale calls based on seven parameters including call duration, minimum frequency, and start frequency, among others. These parameters where used to create criteria to classify the whale calls based on the training data. This method of classification is highly brittle and dependent on training sample and may not be generalizable to different whale species.

Another classifier, the gaussian mixture model (GMM), was used in [2] to classify ship recordings. The authors achieve 75.4% accuracy with this type of classifier.

An overview of machine learning techniques and their applications to the field of acoustics is provided in [7].

B. NEURAL NETWORK BASED APPROACHES

Neural networks offer several advantages over traditional methods. Where traditional methods require expert knowledge and can produce brittle systems, neural networks can learn features without human design. Neural networks do not require expert knowledge and, when provided sufficient training data, can generalize knowledge to previously unseen samples. The requirement for significantly large enough datasets for training is one challenge of utilizing neural networks. This challenge is especially prominent in underwater soundscape where the high costs of obtaining data or sparsity of data lead to small datasets.

1. Introduction to Neural Networks

The concept of neural networks has existed since the 1950s but only in recent years with advances in computational power has their adoption become widespread. The most basic neural networks are feed forward neural networks or multi-layer perceptron (MLP). MLPs are made up of many “neurons” or nodes that take an input multiplied by a weight, a bias, and sum these inputs. The result is then passed through an activation function to produce an output that is sent to the next layer of the network. The weights and biases are the parameters that are “learned” through the training process.

The cost function, or loss function, is used to train the network. The network produces outputs and the error between these outputs and the true labels is determined by the loss function. While many of the parameters of the neural networks will be the same throughout this research, the loss function will be adjusted for multi-class and multi-label tasks.

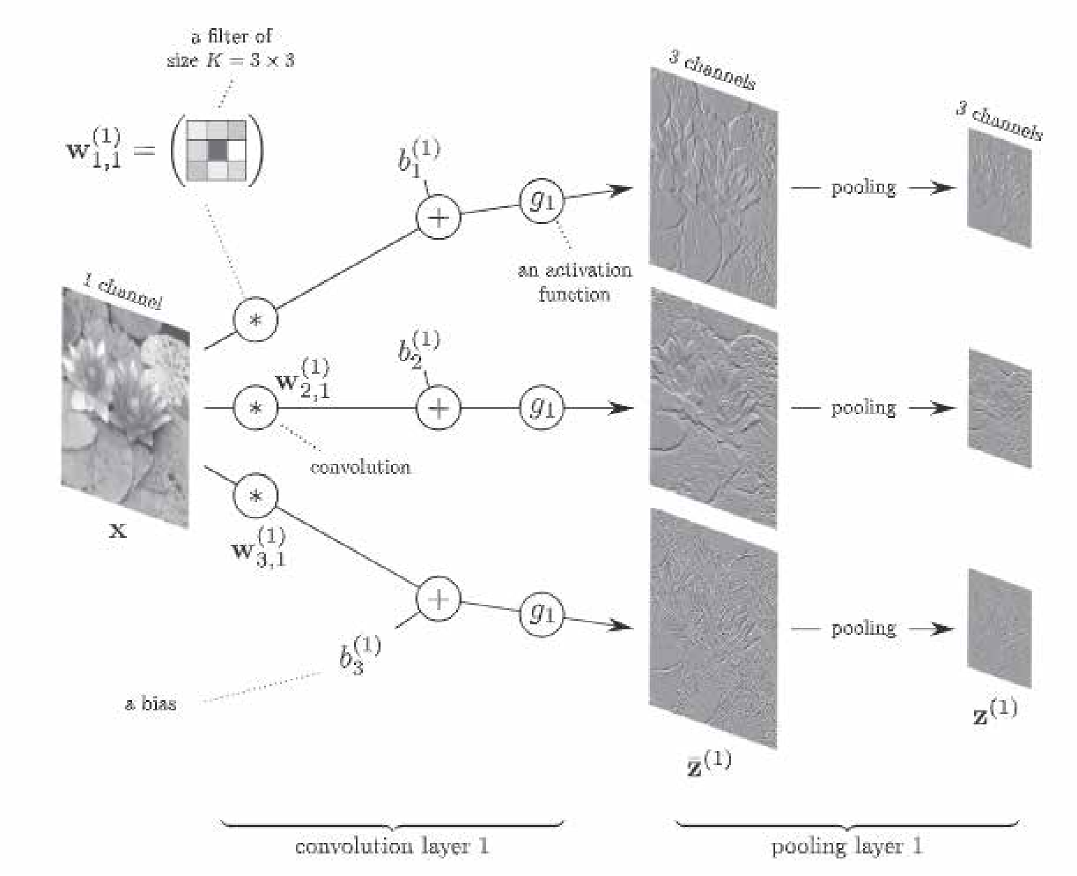


Figure 1. Diagram of Simplified CNN. Source: [7].

2. Convolutional Neural Networks

One type of neural network, the CNN, is especially suited to audio classification tasks. The building block of the CNN is the convolutional layer which consists of one or more filters that are convolved with the input, activating on certain features. A simplified diagram of a CNN is presented in Figure 1, the filter can be seen in the top left corner of the figure. In this figure, three filters are applied to the input image to create three feature maps. These filters have an aperture defined by their shape of height x width; the most common shape is 3x3. This sized filter will have nine weights which are multiplied with the values in the input image and then summed together to create a single output to the next layer. The weights within the filters are learned during training to be activated on certain features in an image like a vertical or horizontal line, or a curve. This filter in then moved across the input in a method called the step or stride. A step of 1 means that the filter is moved one pixel at a time resulting in an output of the same shape as the input. Step sizes greater than one will cause an output smaller than the input to be produced. This is called pooling and is also shown in Figure 1. These smaller level or simpler features (lines or parts of a circle) are built up in higher level filters that can be activated on corners or full circles.

Initial research into soundscape classification using CNNs borrowed from image classification techniques and models. In [8], researchers at Google successfully performed multi-label classification on the audio from millions of YouTube videos. This research used several CNNs developed for images including AlexNet, VGG, Inception, and ResNet, which were adapted to audio data. The most successful model tested was a ResNet 50 model.

Unlike images, there are many different ways to present audio data as input to a model which presents challenges and options to researchers. Audio data can be decomposed into its composite signal frequencies and presented as spectrograms. The generation of spectrograms is discussed in more detail in Chapter III.

Another challenge of building models to classify acoustic data is the number of features present. In an image, each pixel is one input feature to a CNN. Common input architectures for images range from 32x32x3 pixels to 256 x 256x3 resulting in 196608 input features. Audio data from hydrophones is often recorded at high sampling rates, upwards of 40 kHz. Even after down sampling this data, researchers are left with audio sampled at 8 kHz or less. Audio data recorded at 8 kHz can quickly exceed this many input parameters without efforts to reduce the number of parameters. Larger input sizes require more computations and longer training times as well as large datasets to combat model overfitting. Overfitting occurs when a model becomes too adapted to the training dataset and is unable to generalize to new data. This problem is common in models with many parameters to learn over the course of training, including neural networks.

While some methods reduce dimensions by processing the input signals, others target specific frequency bands [9], [10], where only the frequency bands of interest are provided as model input. In [9] the frequency bands of 53–200 Hz and 203–350 Hz are used as input to an SVM and neural network model to determine the range to a passing ship. The use of specific frequency bands is most useful when attempting to classify particular target with a known frequency transmit band such as whale songs or dolphin calls. By targeting the desired frequency band, the amount of noise is reduced in the input data.

Dimensionality reduction, as discussed above, is not the only method used to prevent model overfitting. Two important techniques specific to neural networks, and in this research, are batch normalization and dropout. In dropout, the network, with some specified probability, "drops" or zeros the output of a node in a given network layer [11]. This prevents the network from becoming overly reliant on a few nodes and connections to make inference. The dropout percent is usually set from 20% up to 50%, meaning that up to half of nodes will not be used in a given training batch. The nodes that are dropped changes each training step so that no one node becomes dominate in the network.

Batch normalization is the process of scaling the outputs of a given layer to a mean of 0 and a standard deviation of 1 [12]. This prevents saturation of a node in which the output of one layer becomes so large that it dominates all other inputs to the next layer, preventing other nodes and features from contributing to model output. Batch normalization can be used to regularize the output of every layer in a neural network or just a few. In this research, batch normalization will also be used to normalize the input to the network, scaling all input values to between 0 and 1.

3. CNN Filter Shapes

CNNs that focus on image classification use square kernels of size 3x3 or 5x5, representing the number of pixels they combine at once. While larger kernels can produce better classification results, they can require too many computations to be efficient. Images are treated as orientation invariant; an object upside down is still that object and should be classified as such. Assumptions about image orientation invariance does not transfer to spectrogram images derived from audio. The orientation of spectrograms cannot be changed without changing the meaning of the spectrogram. In contrast to images, the scales of spectrogram axes are usually not of the same scale, with time axes being linear and frequency axes logarithmic. Several studies have explored the use of rectangular kernels in audio classification. Mars et al. use rectangular filters of various sizes to vary the convolution of time and frequency domains. In this study filters of 3x7, 3x1, 11x1, and 1x7 are alternated each layer to alternate how the signal is mixed [13]. This model did not produce better results than a model with all 3x3 kernels. Several studies use rectangular kernels in music classification [14], [15], and [16]. These studies attempt to classify events in music that occur over a wide range of frequencies at the same time or changes in a single frequency band over time.

Another approach to filter design is to create filters specifically designed for the data based on correlation in the spectrograms. In Li et al. [17], custom kernels are generated for the sounds of flapping bird wings and bird songs based on correlation analysis of the sample data. Using this custom kernel approach, Li et al. were able to achieve 5 percent higher accuracy than CNNs using 3x3 or 5x5 kernels. While these kernels produce good results, they cannot be transferred to new data and generating new kernels using this method may be time consuming.

The approach of this research will be to test various kernel shapes and sizes to determine the optimal shape for audio classification. It will be important to balance classification accuracy with computation time required to train a model.

4. Dataset Requirements

One challenge of training neural networks is the high volume of training data that is required to achieve good performance. Neural networks can require millions of training samples. For example, [8] used training sets ranging in size from 23,000 videos to 70 million videos, with performance increasing with increases in training set size. Typically, existing data can be modified in order to boost the number of training samples available. For images this involves rotating, cropping, and altering the white balance, among other methods of data augmentation. While some methods for acoustic data augmentation have been studied [18], most of this work is in the speech or environmental sound domains.

Little work on transferring these methods to the underwater acoustic domain has been accomplished.

In this research, we trialed generating additional data using a GAN to augment the existing dataset. A GAN is a pair of neural networks, one called the generator, the other the discriminator. The generator draws on random distribution as input and outputs synthetic data corresponding to some real distribution, audio from ships in this case. The discriminator takes the generator's output and must determine if it is real or synthetic data. As the generator learns to generate more realistic data, the discriminator must get better at determining real from fake. For more on GANs and the specific GAN used to create the data used in this research see [19].

C. MULTI-CLASS VERSUS MULTI-LABEL

Machine learning systems often make assumptions about how the data will be presented to the classifier and what the expected output will be. For example, in a simple image classifier, the designers may want to classify a picture with a single label such as a "dog" or "cat." This would assume that each image presented to the classifier contains either a dog or cat, not both, and not a bird or deer. While this classifier may perform well with a well curated dataset of dog and cat images, the real world is messier and there are many images with both dogs and cats. Images that contain both classes are not any less "dog" or less "cat" than the images that contain only one. These classes or labels are set at training time and cannot be modified once the model has been trained and deployed.

For single class CNNs, a softmax function is used as the classification layer activation function. The softmax equation for an input vector $\boldsymbol{x}$ is:

$$\sigma(\boldsymbol{x})_i = \frac{e^{\boldsymbol{x}_i}}{\sum_{j=1}^{K} e^{x_j}}$$

where K is the number of classes, and $\boldsymbol{x}_j$ is the output vector. The softmax equation is a generalization of the logistic function. Using the softmax function, each neuron outputs a probability that the input is of a certain class, with all output probabilities summing to 1. The neuron with the highest probability is taken as the networks output classification.

In a multi-label classifier, the designers assume that one or more of the target classes will be present in the examples either individually or together with some overlap. Instead of a binary output of “dog” or “cat” as previously described, the classifier will output a probability associated with each label. For example, an image can have the labels “dog” and “cat” applied to it.

In a multi-label CNN prediction, a probability value between 0 and 1 is assigned per neuron in the output. There are as many output neurons as there are classes. Multiple labels can be predicted when the individual probabilities on the output neurons are greater than the probability threshold which is typically 0.5. The function used as the loss function to train a multi-label network is call binary cross entropy and is:

$$H_p(\mathrm{q}) = \frac{1}{N}\sum_{i=1}^{n} y_i \log\bigl(p(y_i)\bigr) + (1 - y_i)\log\,(1 - p(y_i))$$

where N is the number of samples, y_i is the True label, and $p(y_i)$ is the probability of predicted label.

Labels in a multi-label problem can have a variety of relationships. They can represent a hierarchy with general classes and more specific sub-classes. In this research, multiple labels can be applied to a sample, with all labeled being equal.

III. UNDERWATER SOUNDS

This chapter will focus on sounds in the underwater environment, how they are generated, propagate, and processed, and challenges unique to the underwater environment.

A. UNDERWATER SOUNDS

Sounds in the underwater environment come from many sources including man-made, biological, and environmental. Each of these sources occur in various frequency bands with biological sounds predominate from 8–16 kHz, ships predominantly below 1,000 Hz, and environmental sounds between 100 Hz and 50 kHz [20]. These signals also vary in length, intensity, and frequency distribution. Environmental signals are considered broadband, with no one frequency dominating over others, while man-made signals can be a combination of broadband and narrowband signals, with specific frequencies more detectable than others. Figure 2 shows the distribution of sounds across various frequencies and their sources. The chart shows that there are overlaps in transmission frequencies between various sources and the wide frequency range that sources can emit.

The propagation of underwater sounds is highly dependent on environmental factors including depth of water, temperature gradient, salinity, and bottom sediment composition. Together these factors influence the sound speed profile and how far sounds can travel underwater. Transmission frequency also effects transmission range, with lower frequency sounds traveling much farther than high frequency sounds which are quickly attenuated. These environmental factors and the abundance of sounds can make the underwater acoustic environment more challenging than other acoustic classification tasks like music and urban sounds.

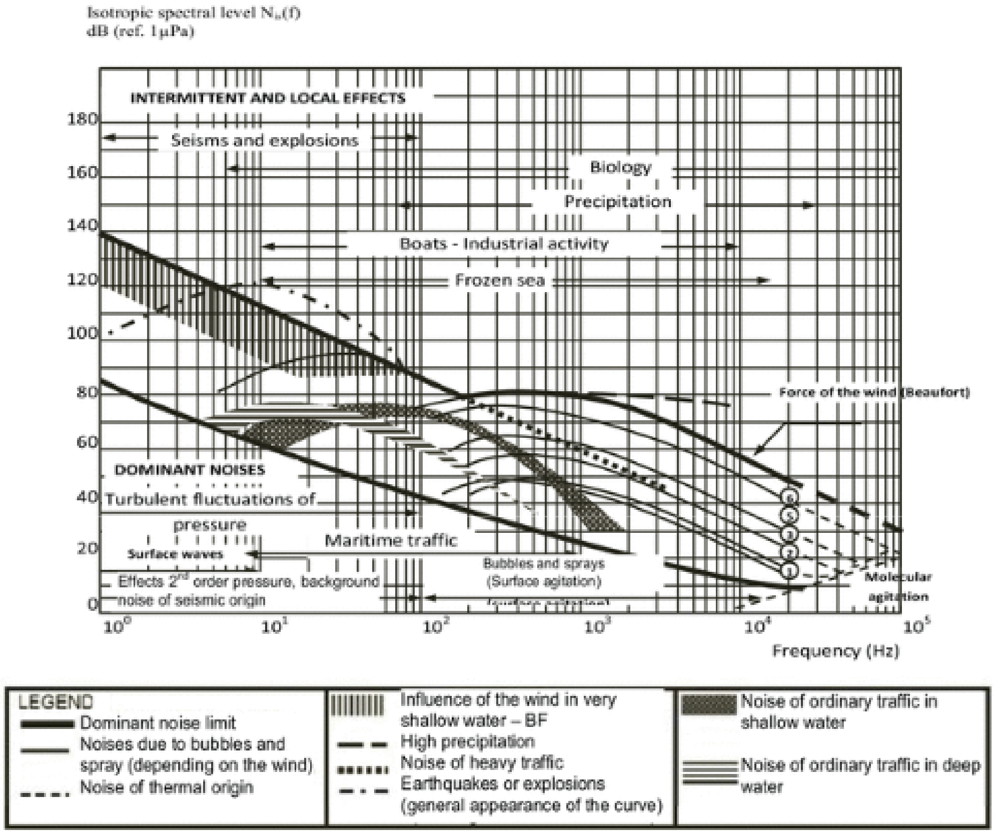


Figure 2. Chart Showing Various Sources of Underwater Sound and Their Frequency. Source: [20].

B. SHIP ACOUSTICS

A ship on the ocean creates acoustic signals from operating machinery, propeller cavitation, and the motion of propeller shafts and reduction gears. Vibration of operating engines and pumps can be transferred through the hull into the water. The size, speed, and aspect to sensor all effect the type and strength of signals received. Arveson and Vendittis provide an overview of the sources and source levels of sound generated by a bulk cargo ship [21]. They show that the main contributors to the ship’s radiated noise signature are the ship’s service diesel generator, main propulsion diesel engine, and propeller blade rate. At low speeds, the diesel engines are the main contributor to noise, however at higher speeds cavitation of the propeller blade becomes the driving force in the ship’s noise. The speed at which this cavitation begins to occur is called cavitation inception or blade rate inception speed. For the ship in [21] this occurred at 10 knots and varies from ship to ship depending on depth of the propellers, fouling of blades and hull, and damage to propeller blades or shaft.

McKenna et al. studied recordings of several commercial ships showing that container ships predominate below 40 Hz and bulk carrier around 100 Hz [22]. They also found that all ships showed asymmetry in their signatures, with bow aspect radiated noise 5–10 dB lower than stern aspects. The relationship between ship size, speed, and emitted sound levels is shown in Figure 3 from [22]. The chart shows the challenges of classification of targets whose sound profile can change with the operating profile of the ship. It also shows that there are differences between ship types that a machine learning algorithm could identify and use to predict the class for previously unseen samples.

The primary task of human sonar operators is to separate ship noise from other noises in the underwater environment so that it can be further investigated and tracked. These ships are tracked either for safety of the ship listening, usually a submarine, or for intelligence collection purposes. Once identified, ships can be further classified by analysis of propeller speed (blade rate), number of propeller blades, and blade pitch. Sonar operators analyze sounds by aural analysis, analysis of specific frequencies emitted by ships, and ship motion relative to the listening location.

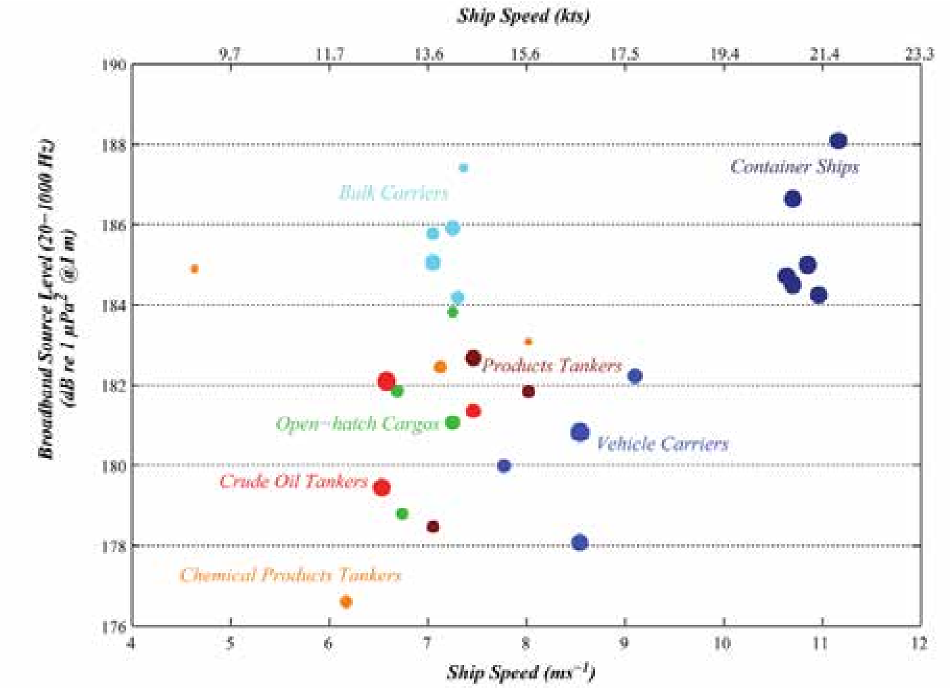


Figure 3. Chart Showing Relationship between Ship Size, Speed, and Radiated Source Level. Source: [22].

C. SPECTROGRAM GENERATION

As discussed in Chapter II, acoustic signals are processed into an image to be used as input to a deep learning classification algorithm. The primary method is to transform the input signal from the time domain to the frequency domain using the short-time Fourier transform (STFT) and display the change in frequency across time. The STFT is used in the analysis of all types of audio signals including speech and environmental sounds. The STFT is the result of the summation of a series of discrete Fourier transforms over a given length of input signal. The Fourier transform decomposes a function into its constituent frequencies. The equation for the STFT of an input signal *x* is given as:

$$X[n,k] = \sum_{m=0}^{L-1} x[m] \cdot w[n-m] e^{-i\frac{2\pi k}{N}m}$$

where *L* is the max number of samples, *w [n-m]* is the window function and *N* is the max frequency.

The spectrogram generated by the STFT can be scaled in several ways to create different representations of the same data. The simplest method is to logarithmically scale the spectrogram. This scaling is performed because sound on the decibel scale is logarithmic. The intensity of the sound in the spectrogram is represented by the color of the image.

Another method is to use a Mel-filter bank to create a mel-log scaled spectrogram, designed to mimic the human hearing scale. The scale of human hearing is not linear, with humans able to discriminate lower frequency sounds better than higher frequency ones. The Mel-filter bank attempts to mimic this scale by combining frequencies in triangular filters. As the frequency becomes higher, more frequencies are combined into one filter. Figure 4 shows the mel spaced filter banks, note that there are more triangular filters in the lower frequencies than in the higher frequencies. Mel frequency scale is linear below 1000 Hz and logarithmic above 1000 Hz. Figure 5 shows a sample spectrogram from the dataset generated using an STFT, on the left, with the Mel-log scaled spectrogram on the right.

While the x axis of the mel-log spectrogram is the same as the log scaled, time in seconds, the y axes change. In the log scaled spectrogram the y axis is frequency from 0 to 2000 Hz in Figure 5, with frequencies binned linearly into the number of FFT bins specified. While the mel-log figure is the number of mel filters chosen. For Figure 5 the y axis is from 0 to 128.

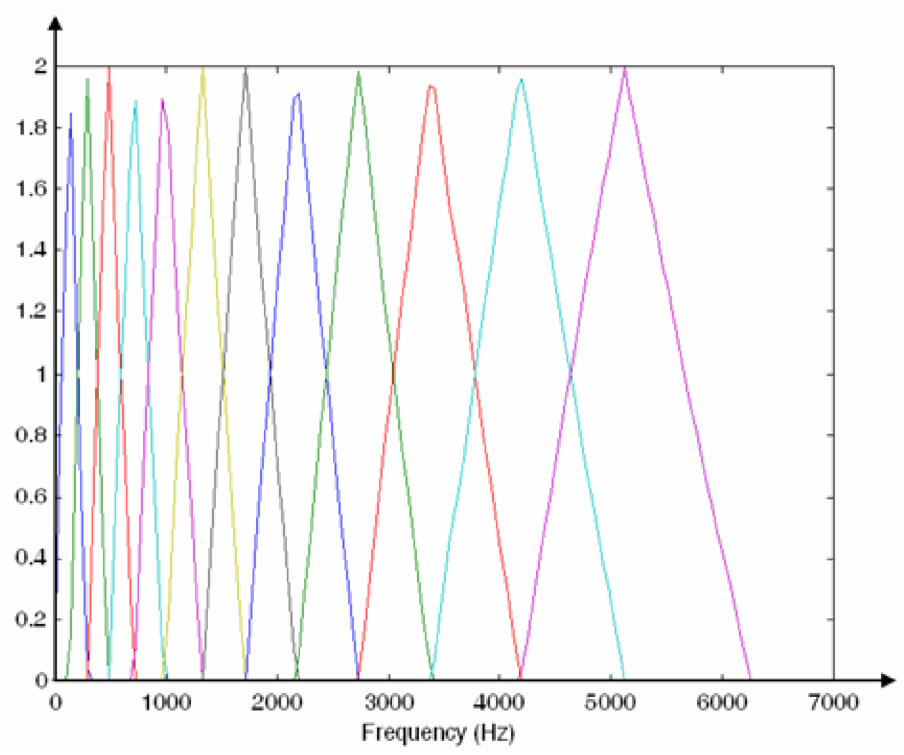
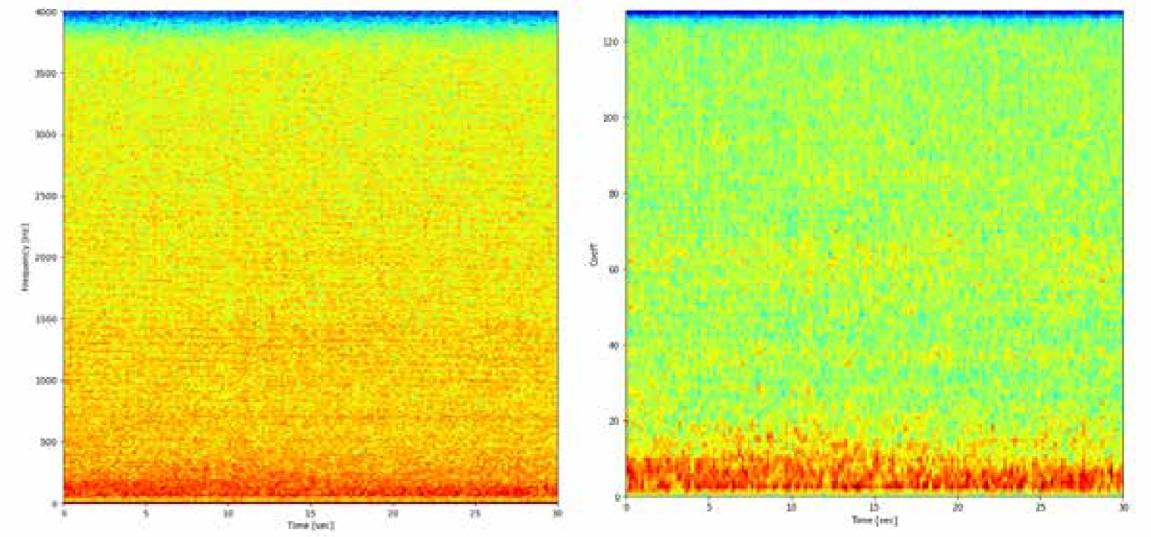


Figure 4. Mel Spaced Filter Bank. Source: [28].



Both spectrograms show a car carrier ship at the closest point of approach at a range less than 1 km. The spectrogram on the left is log scaled STFT, on the right is Mel-Log spectrogram. Pixel color represents sound intensity with red as the highest intensity and blue the lowest.

Figure 5. Example Spectrograms

D. DATASETS

Two datasets are used for training and evaluation of models in this research. The first dataset used was recorded at Thirty Mile Bank off the coast of southern California from December 2012 to April 2013 totaling 2954 hours of recording. The sensor, a High-frequency Acoustic Recording Package (HARP), was deployed in 734 m of water with the sensor 51 m above the sea floor and an original sample rate of 200 kHz. More information on this type of sensor can be found in [23]. This dataset will be referred to as the HARP dataset from here on.

The second dataset was collected in the Monterey Bay from May thru June 2019. This dataset totals 1314 hours and was recorded at a depth of 890 m and an original recording rate of 8 kHz. It was collected using a vector senor which captured not only sound pressure but also data in the x, y, and z axes. Vector sensors offer additionally advantages including the ability to determine range and bearing to the target ship based on inputs from the various axes. This dataset will be referred to as the MARS dataset from here on.

1. Dataset Generation

All audio clips samples were standardized to 30 seconds long and 4 kHz sampling rate. Ship range from the sensor were less than 20 km (10.7 NM) for samples labeled as containing a ship class, time periods where all ships were outside of 30 km are considered “no ship present” samples.

A maximum sample time of 30 seconds was chosen for several reasons. First, the assumption is made that there is little change in the sound emitted by the passing ship during the sample duration, ship noises are stationary within this time frame. Second, an assumption is made that a human operator would need at least 30 seconds of audio or frequency data to detect the presence of the ship. Other studies in audio classification, such as those based on the UrbanSound8k dataset, use shorter time lengths of only a few seconds due to the non-stationary nature of the sounds they are attempting to classify [24].

Input data was preprocessed by first performing a STFT to produce spectrograms, followed by scaling, either log-scaled or Mel-log scaled to serve as input features to the CNN [25].

A time window of 500 msec, equating to 2000 sample points with 1024 FFTs and 75 percent overlap was used for all log-scaled spectrograms. For Mel-log scaled spectrograms 60 and 128 mel bands were tested with 128 bands determined to be optimal.

Several studies in the classification of ship acoustics focus on the lower frequency ranges, and Niu et al. classify on 53–200 Hz and 203–350 Hz frequency ranges while Zak uses 5–200 Hz [28]. Figure 3 shows the relationship between ship type, speed and broadband source levels. While ships of the same type produce higher sound levels at faster speeds, the chart also shows the variation in the source levels of different ship types. From McKenna et al. [22].

2. Dataset Labeling

Ontology is the intersection between human interpretation of the data and the machine interpretation of the data. Ontology drives labeling of data which is one of the main challenges in machine learning. Often training samples have to be hand labeled by humans, a time intensive process that is also prone to error. There is not an open source ontology of ship or underwater sounds. An ontology of sounds is important so that current and future data can be labeled in the same manner, making it easier to transfer machine learning systems and compare test results across studies.

In [2], an ontology of ships is introduced that divides ship types into four classes based on tonnage, a fifth class is used to denote no ship present. Table 1 lists some of the ship types associated with each class. Class A contains small commercial vessels, class B contains pleasure craft and sailboats which are smaller than class A. Class C contains cruise and passenger ships. Class D contains all other ships including container and tanker ships, bulk carriers, and large military ships.

Table 1. Dataset Class List

Class	Ship Designators
A	Fishing Vessel, Tug, Towing Vessel
B	Pleasure Craft, Sailboat, Pilot
C	Passenger ship, Cruise Ship
D	Tanker, Container Ship, Military Ship, Bulk Carrier
E	No ship present, background noise

For the multi-class classification task, samples were only assigned one label indicating which class of the ship was present in the sample. Ship class was determined by matching audio data timestamps with AIS messages and broadcast Maritime Mobile Service Identity (MMSI) numbers or International Maritime Organization (IMO) number where MMSI number could not be found. Both MMSI and IMO numbers allowed for finding precise ship details, such as ship type, length and beam, and dead weight tonnage, in available online ship databases. The IMO number is fixed to a ship, whereas the MMSI number can change with owner or country the vessel is flagged under. All labels were applied automatically from the corresponding AIS data by scraping the ship information from the online ship data website.

For the multi-labeled dataset, samples with more than one ship present were labeled with the class labels of all the ships present at that time within the range of 20 km. Only 11% of the dataset contained samples with more than one ship present at once, which is a common data imbalance in multi-label classification for audio as reported by Google YouTube8M challenge [8].

THIS PAGE INTENTIONALLY LEFT BLANK

IV. METHODOLOGY

This section describes the custom CNN architecture and the tests and hyperparameter searches that lead to the chosen architecture.

A. CUSTOM CNN ARCHITECTURE

In the process of evaluating existing models several choices in model design became apparent. How many layers to have, the number of filters in each layer, and the size of the filters themselves. Several kernel size ratios were tested, of 1:1, 2:1, 3:1, and 4:1, with 2:1 being optimal based on test set accuracy and 4:1 performing the worst. Results of these tests are shown in Figure 6. Filter shapes of 10x5 and 5x10 were trialed on several tests during the design of the model. There was no detectable difference between these filter sizes.

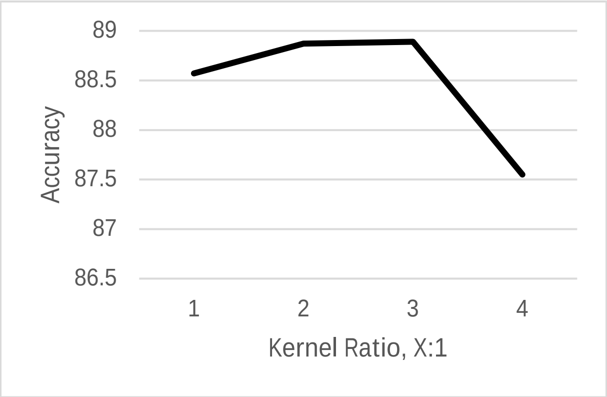


Chart showing the accuracy of various kernel size ratios. The chart shows increased performance with 2:1 and 3:1 ratios with significant drop at 4:1.

Figure 6. Classifier Accuracy with Respect to Kernel Ratio

Additional tests were performed to determine the number of convolutional blocks to include in the model. Tests with five blocks had an average accuracy of 84.36 % while models with three blocks had an average accuracy of 73.58%. These tests indicated that five convolutional blocks were an optimal choice. The addition of any more blocks would have reduced the output too far because the output of each block is half the size of its input. The addition of any more blocks would require a redesign of the individual block structure.

In the proposed CNN, shown in Figure 7, kernels of 10x5 were used to vary the convolution of time and frequency domains. These kernel sizes were fixed throughout every layer of the network. An initial Batch Normalization layer was used to normalize input spectrograms. Based on [26] it was determined that increasing the number of filters throughout the network was desirable. The hypothesis behind this design is that there are a few low-level features to be learned while there are more high-level features that many higher-level filters learn. The initial layers contained 16 filters with 16 added in each additional set. After each block of two convolutional layers, the input size to the next block is cut in half by a max pooling layer with a stride of 2 by 2.

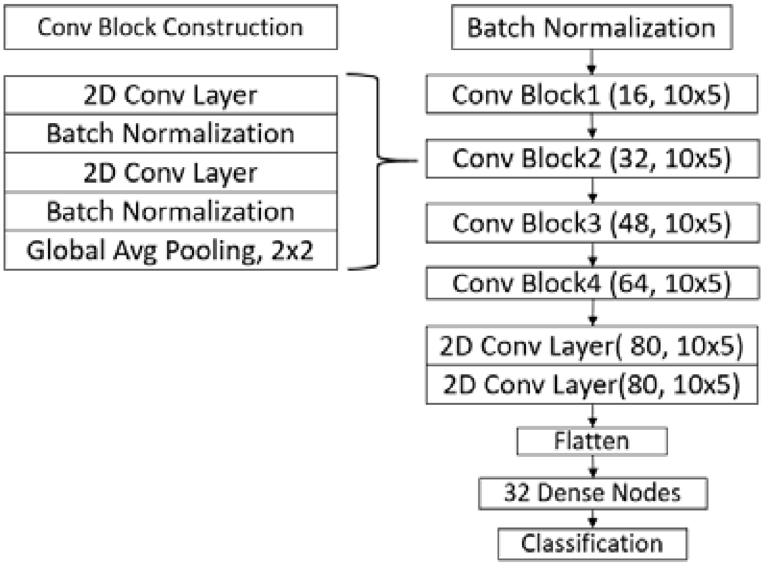


Figure 7. Diagram of Proposed Custom CNN Architecture

Each convolutional layer was followed by a batch normalization layer and ReLU activation. The classification head of the network was comprised of 32 dense nodes with ReLU activation, a dropout layer of 0.2 for regularization, and finally, the classification layer with softmax activation function. When performing multi-label classification, the sigmoid activation function was used on each neuron in the classification layer and binary cross entropy was used as the network loss function.

B. EXPERIMENT METHODOLOGY

All experiments described divided the dataset used into training, testing, and validation sets. The training set was 80% of the entire dataset with testing and validation each 10%. Training times were 50 epochs for all tests, with tests of 150 and 200 epochs confirming that model performance plateaued after 50 epochs. Since there are no benchmark datasets for underwater soundscapes, a comparison was needed to evaluate the performance of the custom CNN. The proposed CNN architecture was evaluated against a the ResNet44 v1 architecture trained on the dataset and several popular models pretrained on the ImageNet dataset. The ResNet44 model is based on [30] and so call because it has 44 convolutional layers. While the custom model uses dense node layers prior to the classification layer and after the convolutional layers, the ResNet44 models utilizes a global average pooling layer instead.

The pretrained models used were the VGG, MobileNet v2, and InceptionNet architectures. All of these are available via the Tensorflow applications libraries. Only the classification layer of the pre-trained models were retrained on the dataset for testing. The pretrained models were designed for images and therefore have three channels instead of a single channel used by the custom and ResNet44 models. To use these models, the input data was replicated so that each of the three channels had the same input data.

The MARS sensor collects data in four channels as described in Chapter III. While some experiments only use the pressure (sound) channel of this dataset, one set of experiments used all four channels in parallel. This experiment was to determine if the x, y, and z channels provide additional data to the classifier that improve its performance.

The model was trained using the Adam optimizer and a learning rate of 0.0001 for all tests [25]. Dataset sizes varied for each test; dataset sizes are provided along with the test results. For each test, the dataset was divided into three sets, with 80 percent used for training, 10 percent for validation, and a 10 percent holdout set used for test evaluation.

The model was coded using Tensorflow v2.1and all tests were performed on an NVIDIA RTX 8000 graphics processing units.

C. GAN DATA GENERATION

The ability of GANs to serve as useful generators of synthetic data was trialed. A subset of samples taken from the larger HARP dataset was used as input to the GAN. These samples were ships that passed the sensor at a distance of 4 km or closer. This was done in order to provide the best examples for the GAN to learn from. Several sample classes were evaluated, first large ships and small ships were tested to determine if it was possible for the GAN to generate ship like sounds. Next samples from container ships, tugs, and warships (a destroyer) were evaluated to determine if the GAN could generate more specific types of ship sounds. The GAN generated samples that were 16384 samples long each, creating 4 second samples at 4 kHz.

Tests were conducted with training conducted by training a classifier model on only the real data or on a mix of real and synthetic data. When the classifier was trained only on real data, it was evaluated on the synthetic data to test if the GAN generated realistic data. When the classifier was trained on a mix of real and synthetic data, the classifier was tested on a holdout set of only real data to determine if the addition of synthetic data improved classifier performance when compared to one trained on real data alone.

A third test was conducted to determine if the addition of synthetic data to a single class would boost the performance of that class. This scenario is very common in machine learning problems, where one class is under-represented in the training dataset due to a lack of available data for that particular class. Here, synthetic data was generated for the "warship" class and added to real training data for a container ship class, tug class and warship class. The classifier was then tested on only real data from these three classes. These results are compared to a classifier trained only on real data from the three classes.

V. RESULTS

This section presents the results of tests run on both and custom CNN model and other models used for comparison. Results are presented for multi-class, multi-label, and classification with synthetic data. In addition to numerical results, some plots of inference on target ships are shown to demonstrate the challenges of real-world applications.

A. MULTI CLASS

1. HARP Dataset

Overall, mel-log spectrograms resulted in higher classifier accuracy than log scaled spectrograms, for the multi-class task as shown in Table 2. These results are similar to [13] where mel-log spectrograms showed better performance than other methods chosen for the multi-class task in urban soundscape classification. For these reasons, only mel-log spectrograms were used during the multi-label task, shown in Table 3.

In the multi-class task, the custom model performed better than the ResNet44 v1 model by achieving higher average F1 score by 6%. The pretrained models could have achieved higher scores if some of the features were trained with fine tuning, however only the classification layer was trained. Additionally, the pretrained models suffered more from class imbalances than either model trained from scratch on the dataset.

Table 2. Table 2: HARP Multi-class Results

	Fully Trained				Pretrained					
F1 Score	Custom		RESNET		VGG		INCEPTION		MOBILENET	
Input Type	STFT	MEL	STFT	MEL	STFT	MEL	STFT	MEL	STFT	MEL
Class A	0.82	0.84	0.69	0.77	0.45	0.55	0.57	0.60	0.60	0.59
Class B	0.72	0.79	0.47	0.71	0.02	0.32	0.43	0.42	0.40	0.42
Class C	0.84	0.89	0.77	0.83	0.28	0.38	0.59	0.55	0.56	0.54
Class D	0.88	0.92	0.80	0.87	0.68	0.70	0.75	0.74	0.75	0.73
Class E	0.93	0.98	0.87	0.94	0.71	0.79	0.82	0.80	0.81	0.81
Average	0.838	**0.884**	0.72	0.824	0.428	0.548	0.632	0.622	0.624	0.618

An example of classification of the single target ship is shown in Figure 8. In the figure, green points represent classification of “no ship present,” “clssE,” while blue points represent correct classification of “classD.” In this example, the classifier is unable to properly classify target ship as it approaches the sensor due to the bow null of the ship blocking the sound of the propeller and engines until approximately8 km. After the ship passes the sensor, the classifier is able to correctly classify the target ship to approximately 25 km in range, beyond the 20 km training range.

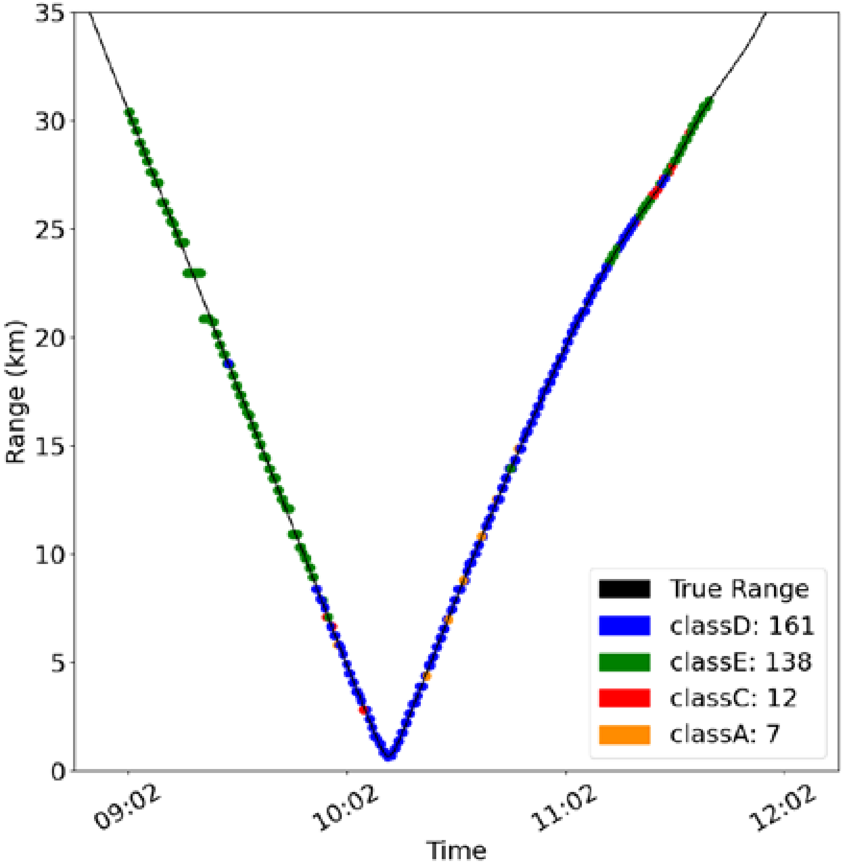


Figure 8. Time-Range Plot of Passing Ship

2. MARS Dataset

Due to its location, the distribution of ship types in the MARS dataset is significantly different than the HARP dataset. There are far fewer class D large ships in the

MARS dataset along with a larger set of small vessels in classes A and B. This difference in distribution allowed for different analysis of the MARS data than the HARP data.

In addition to broad classes, specific ship types were also analyzed for the ability of the model to classify, results of this analysis are presented in Table 3. Overall, the custom model performed better than the ResNet44 model by 12.6 % on average. Additionally, the custom model was less susceptible to variation in class sizes. This test demonstrates the ability of CNNs to predict specific ship types, not just broad categories, if given a well curated training dataset.

Table 3. MARS Multi-class F1 Score Results

Label	Custom Model	ResNet44	Num Samples in Test Set
Fishing Vessel	0.73	0.64	1500
Container Ship	0.64	0.28	278
Offshore Tug	0.76	0.72	721
Pleasure Craft	0.73	0.66	1474
Sailboat	0.73	0.62	1486
Average	0.718	0.592	

3. Vector Sensor

Results in Table 4 shown the performance of the custom and ResNet44 models on multi-channel data from the MARS dataset. These results are compared to single channel (pressure channel only) results for the same dataset and train, test split. The custom model performs better than the ResNet44 model, this is similar to pervious results on the HARP dataset. Performance is better with four channels than 1 by an average of 6 percent for the ResNet44 model and 2 percent for the custom model. Lower overall performance in this test may be due to differences in the MARS dataset versus the HARP dataset. The results also show that additional channels provide some additional benefit for classification as F1 and accuracy scores for the multi-channel model were higher than single channel for both models tested.

Table 4. F1 Scores for Multi and Single Channel Data

	ResNet44		Custom Model		
	4 Channels	1 Channel	4 Channels	1 Channel	Test set size
Class A	0.59	0.56	0.64	0.65	5410
Class B	0.52	0.47	0.55	0.55	2653
Class C	0.65	0.63	0.69	0.66	6535
Class D	0.40	0.31	0.52	0.47	406
Class E	0.75	0.67	0.79	0.75	739
Average	0.582	0.528	0.638	0.616	-

B. MULTI-LABEL

Multi-label classification was performed exclusively on the HARP dataset. The breakdown of the multi-label dataset is shown in Table 5. Only 10 percent of the total dataset had samples with more than one label.

Table 5. Multi-label Dataset Label Distribution

Label Type	Count	Percent of Total Dataset
Single Label	45346	89.6
Two Labels	4879	9.6
Three Labels	373	0.74
Four Labels	2	0.004

For the multi-label classification task, evaluation metrics of area under the curve (AUC) of the precision recall curve and micro-F1 score where chosen due to the imbalanced nature of the dataset. These metrics were also used in the 2019 Detection and Classification of Acoustics Scenes and Events in which urban soundscapes were used in a multi-label classification challenge [29]. Results of the multi-label test, Table 6, show that the custom model performs well in the multi-label classification task outperforming or comparable to the multi-class case.

Table 6. Multi-label Results

	Custom	ResNet 44 v1
Avg. micro-F1	0.97	0.90
AUCPR	0.896	0.730

An example of the results for multi-label classification is shown in Figure 9. In this example, multiple ships overlap within the range which they can be classified. The classifier preferences the ship that is closest to the sensor in the sample. When the tug, designated by the blue points, reaches its closets range from the sensor, the classifier also predicts the "classD" labels, the largest class of ships. This is most likely due to the high volume of energy in the water as the tug passes the sensor at a close range.

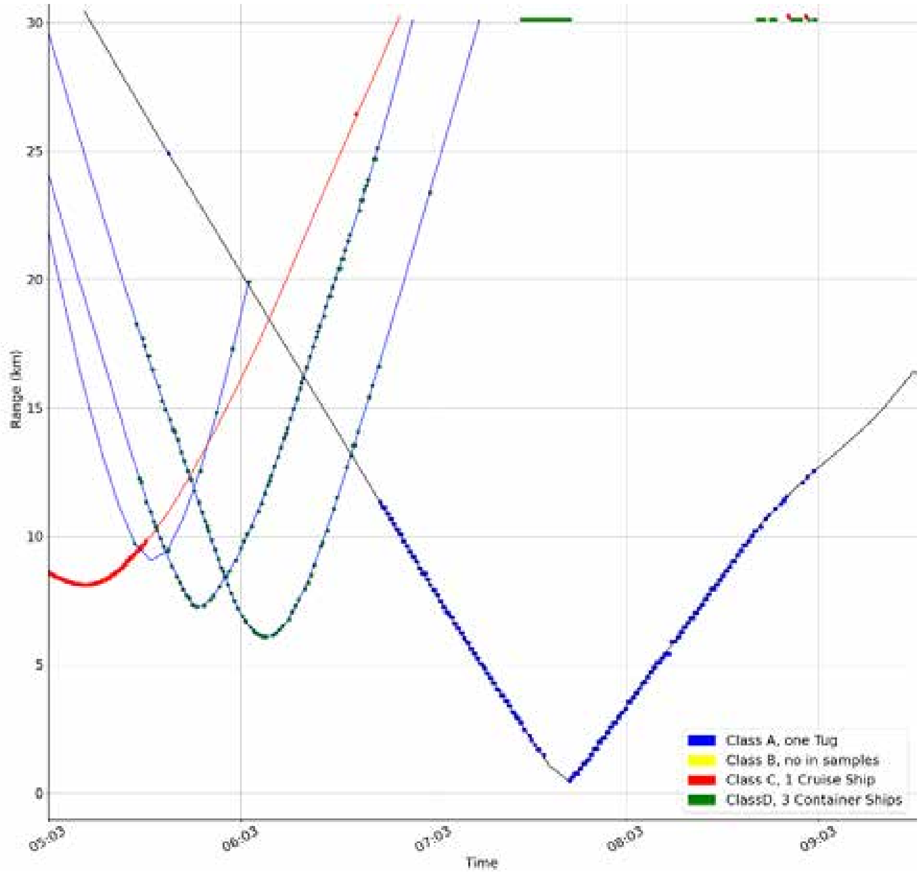


Figure 9. Time-Range Plot Showing Classification when Using Multi-Label Classification

C. DATA AUGMENTATION WITH GANS

1. Evaluation of Synthetic Data Realness

In the first test, the classifier was trained on real data only and evaluated on synthetic data. For each class, 4000 samples were used to train the model, results are shown in Table 7. Both the ResNet44 and custom models were able to correctly classify the synthetic data at the same or better levels than the real data test sets. This indicates that the GAN was able to produce realistic data. It was able to learn the defining characteristics of ship audio and emphasize those features which is desirable for data augmentation. Based on the higher accuracy of the mel-log spectrograms and the custom model from this test, only mel-log spectrograms and the custom model were used in further testing.

Table 7. Accuracy of Classifier When Trained on Real Data Only

		RESNET 44		*CUSTOM MODEL*	
Data Source	CLASS	STFT	MEL LOG	STFT	MEL LOG
GAN	largeShip	0.794	0.973	0.977	0.988
	smallShip	0.981	0.979	0.986	0.986
	OVERALL	0.888	0.976	0.982	0.988
Real Data	largeShip	0.953	0.977	0.973	0.982
	smallShip	0.929	0.944	0.961	0.977
	Overall	0.942	0.963	0.968	0.98

2. Augmenting with Synthetic Data

Tests were conducted with increasing amounts of GAN generated synthetic data added to the training dataset. Table 8 shows the results for both real data training only and three additional amounts of synthetic data added to the real data. Classifier performance went down slightly with the addition of 1000 and 5000 synthetic samples but up with only 2000 samples. This indicates that there is an ideal ration of real to synthetic data in the training set. Too much or not enough synthetic data with adversely affect the classifier’s performance on real test data.

Table 8. F1 Scores for Training with Synthetic Samples

	Real Data	1K Synthetic Samples	2K Synthetic Samples	5K Synthetic Samples
largeShip	0.99	0.98	0.98	0.97
smallShip	0.98	0.97	0.97	0.96
Overall	0.98	0.977	0.98	0.97

3. Boosting Single Class Performance

The results for adding synthetic data of only the “warship” class are shown in Table 9. The custom model was trained on only real data, and then two training sets with 2,000 synthetic samples and 10,000 synthetic samples added. Results from the previous experiment indicated that there is an optimal amount of additional synthetic training data that can be added, which is confirmed to some degree here. Minimal change was observed with the addition of just 2,000 samples, however performance for all classes decreased with the addition of 10,000 samples. It is also possible that while the GAN was good at generating generic ship sounds in the previous tests, it was unable to generate more specific ship sounds here. More testing is required to determine the exact cause of the decrease in performance.

Table 9. F1 Scores for Custom Model for Single Class Boosting

	Real Data Only	2K Synthetic Samples	10K Synthetic Samples	Number of Real Samples
Container Ship	0.92	0.92	0.90	38,974
Tug	0.95	0.94	0.93	51,556
Warship	0.95	0.96	0.89	7,417
Overall	0.94	0.94	0.90	-

THIS PAGE INTENTIONALLY LEFT BLANK

VI. CONCLUSIONS AND FUTURE WORK

A. CONCLUSION

This research explored several topics in underwater soundscape classification including multi-class and multi-label classification, and dataset augmentation with synthetic data. Each of these areas is important in and of itself; however, a system level approach as applied in this thesis provides an overview of a staggered impact that each of these contributions provide.

This research demonstrated the ability of CNNs to perform multi-label classification of ship acoustics in the underwater environment. Both the custom and ResNet44 models were able to classify multi-label samples with an average micro-F1 score of greater than 90%. The custom model achieved better performance than the ResNet44 model by 7%.

The same models also demonstrated the ability to perform multi-class classification on broad categories of ship types and on specific ship designations. The custom model was successful in classifying the multi-class dataset with an average F1 score of 88.4%, better than all other models tested.

The custom CNN model's superior performance in both tasks is because it takes advantage of domain specific knowledge to create filters that combine time and frequency domain data in rectangular filters. The model's performance relative to the other selected models demonstrates the need for the design of acoustic specific models, not just the adaptation of architectures built for image-based applications.

This research also demonstrated the ability for GANs to generate realistic ship like sounds and for those sounds to be incorporated into a classifier's training dataset. The incorporation of this data did show that it could improve the performance of a classifier when operating on real world data. However, that performance is dependent on the amount of synthetic data included in the training dataset. While the GAN tested was able to generate generic ship sounds, it was unable to generate more specific sounds. This may be due to the GAN or the training data supplied to the GAN. More research on the generation of specific ships sounds is required to make a definitive ruling. One limitation of this series of tests is that the data used to train the GAN and classifier is the same data and could in effect just create copies of the original data, by sampling from a very narrow distribution of data representations. This decision was made as this test was a proof of concept that this methodology works. This limitation could be overcome by gathering larger datasets to train GAN to promote better diversity in generated sounds.

While the success of both multi-class and multi-label classification was demonstrated in this research, it has also highlighted some of the challenges. The presented examples show that both environmental and ship aspect challenges of classification of underwater acoustic data.

Despite these challenges, there are many exciting potential applications that this research enables. Automatic detection and classification of ships by underwater sonar sensors, whether on board ships, submarines, or fixed undersea arrays, is the most obvious application. Many naval platforms rely on human operators to monitor sonar systems for threat contacts, a laborious and tedious task. The machine learning methods demonstrated in this research could be used assist human operators in their tasks of locating ships. With enough maturity, these models could eventually replace some of these human operators. These models could also be deployed on UUVs to provide the ability to detect and track ships and remotely monitor areas of the ocean for enemy warships. The ability to detect and respond to other ships in the ocean is a capability that does not currently exist in any UUV. We hope this research is only the beginning of more advanced study in this area.

B. FUTURE WORK

As discussed in the previous section, this research demonstrated the advantages of designing and training models specifically for acoustics. There is still more work that can be accomplished with respect to studying model architectures and hyperparameters. Future work should consider potential improvements that can be made to the custom model presented in this work. In the future, the need exists to expand the current work to include multi-instance multi-label classification task which would capture the presence of more than one ship of the same type in a sample of audio data.

Future work in this area should include an exploration of environmental effects on classifier performance. If a classifier is trained on a dataset collected from location A, it is not known how well it performs on data collected at a location with very different ocean conditions. The collection of more datasets from a variety of locations will assist in this work. Understanding how models are affected by changing environmental conditions is extremely important to the Navy, as ships are deployed in far different environments than where they conduct training operations near their home port.

Another area for research is the best methods of data augmentation for underwater sounds. Data augmentation is an important step in expanding existing datasets and making them more robust to unseen examples. While some methods exist for data augmentation of human speech, these methods may not be transferable to underwater sounds and more exploration is needed. While some research on the inclusion of synthetic data was explored, no research was conducted on altering the real data. This could be done by shifting or time stretching the samples, or by adding other sounds, like whale songs, to samples of target ships.

The research on ship labels can also be extended. Research into an ontology for underwater sounds in needed to be able to produce better labeling of gathered data. Additionally, while this research demonstrated broad and specific labeling of ships, future work could combine these into a label hierarchy.

Only one experiment was conducted with all four channels of the MARS dataset collected with a vector sensor. Since these sensors collect additional data, they provide an opportunity to attempt regression inference. Models that predict range to the ship or bearing from the sensor could take advantage of the additional channels to learn this data.

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[2] D. Santos-Domínguez, S. Torres-Guijarro, A. Cardenal-López, and A. Pena-Gimenez, “ShipsEar: An underwater vessel noise database,” *Applied Acoustics*, vol. 113, 2016, pp. 64–69.

[3] H. Niu, E. Ozanich, and P. Gerstoft, “Ship localization in Santa Barbara Channel using machine learning classifiers” *Journal of the Acoustical Society of America* vol. 142, Nov. 2017.

[4] B. Werner, “Navy Awards Boeing $43 Million to Build Four Orca XLUUVs” *USNI News*, Apr 17, 2019. [Online]. https://news.usni.org/2020/01/31/navy-industry-pursuing-autonomy-software-reliable-hme-systems-for-unmanned-ships

[5] J. Stanford, “Keyport is home to Navy’s first unmanned undersea vehicle squadron,” July 3, 2018. [Online]. https://www.kitsapsun.com/story/news/local/navy/2018/07/03/keyport-home-navy-first-underwater-drone-squadron/751928002/

[4] L. Fillinger, A. J. Hunter, M. C. C. M. Zampolli, and M. C. Clarijs, “Passive acoustic detection of closed-circuit underwater breathing apparatus in an operational port environment,” *The Journal of the Acoustical Society of America*, vol. 132, 2012 pp. 310–316.

[6] D. Gillespie, “Detection and classification of right whale calls using an ‘edge’ detector operating on a smoothed spectrogram,” *Canadian Acoustics*, 2004

[7] M. J. Bianco, P. Gerstoft, J. Traer, E. Ozanich, M. A. Roch, S. Gannot, and C. A. Deledalle. “Machine learning in acoustics: theory and applications.” *Journal of the Acoustical Society of America*, vol. 146, no. 5, 2019, pp. 3590–3628.

[8] S. Hershey et al. “CNN architectures for large-scale audio classification,” *International Conference on Acoustics, Speech and Signal Processing*, 2017.

[9] C. McQuay, F. Sattar, and P. Driessen, “Deep learning for hydrophone big data,” *2017 IEEE Pacific Rim Conference on Communications, Computers, and Signal Processing (PACRIM)*, 2017, pp. 1–6.

[10] A. Tesei, R. Been, and F. Meyer. “Continuous real-time acoustic surveillance of fast surface vessels.” *Procs. of UACE’17*. 2017.

[11] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, “Improving neural networks by preventing co-adaption of feature detectors” ArXiv Preprint ArXiv: 1207.0580, 2012

[12] S. Ioffe, C. Szegedy. “Batch normalization: accelerating deep neural network training by reducing internal covariate shift.” *Proceedings of the 32nd International Conference on Machine Learning*, 2015, pp. 448–456

[13] Mars, Rohith et al. “Acoustic scene classification from binaural signals using convolutional neural networks.” *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, 2019.

[14] Pons, Jordi et al. “Experimenting with musically motivated convolutional neural networks.” *2016 14th International Workshop on Content-Based Multimedia Indexing (CBMI)*, 2016, pp. 1–6.

[15] Pons, Jordi, and X. Serra. “Designing efficient architectures for modeling temporal features with convolutional neural networks.” *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 2472–2476.

[16] Pons, Jordi et al. “Timbre analysis of music audio signals with convolutional neural networks.” *25th European Signal Processing Conference (EUSIPCO),* 2017, pp. 2744–2748.

[17] Li, Xingyi et al. “Filter shaping for convolutional neural networks.” *ICLR 2017 : International Conference on Learning Representations*, 2017, 2017.

[18] Nanni, Loris et al. “An ensemble of convolutional neural networks for audio classification.” ArXiv Preprint ArXiv:2007.07966, 2020.

[19] N. Thiem, “Creating underwater sounds using generative adversarial networks,” M. S. thesis, Dept of Computer Science, NPS, Monterey, CA, USA, Sep 2020

[20] J. Marage and Y. Mori, *Sonar and Underwater Acoustics*. Hoboken, NJ, USA, 2013.

[21] P. T. Arveson, and D. J. Vendittis. “Radiated noise characteristics of a modern cargo ship.” *Journal of the Acoustical Society of America*, vol. 107, no. 1, 2000, pp. 118–129.

[22] M. F. McKenna et al. “Underwater radiated noise from modern commercial ships.” *Journal of the Acoustical Society of America*, vol. 131, no. 1, 2012, pp. 92–103.

[23] S. M. Wiggins, and J. A. Hildebrand. “High-frequency acoustic recording package (HARP) for broad-band, long-term marine mammal monitoring,” *Symposium on Underwater Technology and Workshop on Scientific Use of Submarine Cables and Related Technologies*, 17–20 Apr 2007, IEEE, Tokyo, p 551–557.

[24] J. Salamon et al. “A dataset and taxonomy for urban sound research.” *Proceedings of the 22nd ACM International Conference on Multimedia*, 2014, pp. 1041–1044.

[25] M. Huzaifah, “Comparison of time-frequency representations for environmental sound classification using convolutional neural networks.” ArXiv Preprint ArXiv:1706.07156, 2017.

[26] Ozyildirim, B. Melis, and S. Kartal. “Comparison of deep convolutional neural network structures the effect of layer counts and kernel sizes.” *Fourth International Conference on Advances in Information Processing and Communication Technology - IPCT* 2016, 2016, pp. 16–19.

[27] Kingma, Diederik P., and Jimmy Lei Ba. “Adam: a method for stochastic optimization.” *ICLR 2015: International Conference on Learning Representations* 2015, 2015.

[28] A. Zak, “Ship’s hydroacoustics signatures classification using neural networks.” 2011. IntechOpen, London, UK. [Online]. Available: https://doi.org/10.5772/14016

[29] D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, M. Lagrange and M. D. Plumbley, “Detection and classification of acoustic scenes and events: An IEEE AASP challenge,” *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, 2013, pp. 1–4, doi: 10.1109/WASPAA.2013.6701819

[30] He, Kaiming et al. “Deep residual learning for image recognition.” *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

THIS PAGE INTENTIONALLY LEFT BLANK

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California